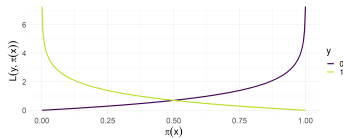


Introduction to Machine Learning

Logistic regression (Deep-Dive)



Learning goals

- Derive the gradient of the logistic regression
- Derive the Hessian of the logistic regression
- Show that the logistic regression is a convex problem

LOGISTIC REGRESSION: RISK PROBLEM

Given $n \in \mathbb{N}$ observations $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{0, 1\}$ we want to minimize the following risk

$$\mathcal{R}_{\text{emp}} = - \sum_{i=1}^n y^{(i)} \log \left(\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right) + \left(1 - y^{(i)} \log(1 - \pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right)) \right)$$

with respect to $\boldsymbol{\theta}$ where the probabilistic classifier

$$\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) = s \left(f \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right),$$

the sigmoid function $s(f) = \frac{1}{1 + \exp(-f)}$ and the score $f \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) = \boldsymbol{\theta}^\top \mathbf{x}$.

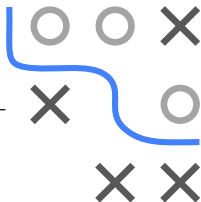
NB: Note that $\frac{\partial}{\partial f} s(f) = s(f)(1 - s(f))$ and $\frac{\partial f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\mathbf{x}^{(i)})^\top$.



LOGISTIC REGRESSION: GRADIENT

We find the gradient of logistic regression with the chain rule, s.t.,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} &= - \sum_{i=1}^n \frac{\partial}{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} y^{(i)} \log(\pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})) \frac{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \\ &\quad \frac{\partial}{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} (1 - y^{(i)}) \log(1 - \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})) \frac{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= - \sum_{i=1}^n \frac{y^{(i)}}{\pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} \frac{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{1 - y^{(i)}}{1 - \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} \frac{\partial \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= - \sum_{i=1}^n \left(\frac{y^{(i)}}{\pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} - \frac{1 - y^{(i)}}{1 - \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})} \right) \frac{\partial s(f(\mathbf{x}^{(i)} | \boldsymbol{\theta}))}{\partial f(\mathbf{x}^{(i)} | \boldsymbol{\theta})} \frac{\partial f(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= - \sum_{i=1}^n \left(y^{(i)} (1 - \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})) - (1 - y^{(i)}) \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \right) (\mathbf{x}^{(i)})^\top.\end{aligned}$$



LOGISTIC REGRESSION: GRADIENT

$$= \sum_{i=1}^n \left(\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) - y^{(i)} \right) \left(\mathbf{x}^{(i)} \right)^\top.$$

$$= \left(\pi(\mathbf{X} \mid \boldsymbol{\theta}) - \mathbf{y} \right)^\top \mathbf{X}$$



where $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})^\top \in \mathbb{R}^{n \times d}$, $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})^\top$, $\pi(\mathbf{X} | \boldsymbol{\theta}) = (\pi(\mathbf{x}^{(1)} | \boldsymbol{\theta}), \dots, \pi(\mathbf{x}^{(n)} | \boldsymbol{\theta}))^\top \in \mathbb{R}^n$.

$$\Rightarrow \text{The gradient } \nabla_{\theta} \mathcal{R}_{\text{emp}} = \left(\frac{\partial}{\partial \theta} \mathcal{R}_{\text{emp}} \right)^{\top} = \mathbf{X}^{\top} (\pi(\mathbf{X} | \theta) - \mathbf{y})$$

This formula can now be used in gradient descent and its friends.

LOGISTIC REGRESSION: HESSIAN

We find the Hessian via differentiation, s.t.,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}^2 \mathcal{R}_{\text{emp}} &= \frac{\partial^2}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \frac{\partial}{\partial \boldsymbol{\theta}^\top} \sum_{i=1}^n \left(\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) - y^{(i)} \right) \left(\mathbf{x}^{(i)} \right)^\top \\ &= \sum_{i=1}^n \mathbf{x}^{(i)} \left(\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \left(1 - \pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right) \right) \left(\mathbf{x}^{(i)} \right)^\top \\ &= \mathbf{X}^\top \mathbf{D} \mathbf{X}\end{aligned}$$



where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal

$$\left(\pi \left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta} \right) \left(1 - \pi \left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta} \right) \right), \dots, \pi \left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta} \right) \left(1 - \pi \left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta} \right) \right) \right).$$

Can now be used in Newton-Raphson and other 2nd order optimizers.

LOGISTIC REGRESSION: CONVEXITY

Finally, we check that logistic regression is a convex problem:

We define the diagonal matrix $\bar{\mathbf{D}} \in \mathbb{R}^{n \times n}$ with diagonal

$$\left(\sqrt{\pi(\mathbf{x}^{(1)} | \boldsymbol{\theta})(1 - \pi(\mathbf{x}^{(1)} | \boldsymbol{\theta}))}, \dots, \sqrt{\pi(\mathbf{x}^{(n)} | \boldsymbol{\theta})(1 - \pi(\mathbf{x}^{(n)} | \boldsymbol{\theta}))} \right)$$

which is possible since π maps into $(0, 1)$.

With this, we get for any $\mathbf{w} \in \mathbb{R}^d$ that

$$\mathbf{w}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{R}_{\text{emp}} \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \bar{\mathbf{D}}^\top \bar{\mathbf{D}} \mathbf{X} \mathbf{w} = (\bar{\mathbf{D}} \mathbf{X} \mathbf{w})^\top \bar{\mathbf{D}} \mathbf{X} \mathbf{w} = \|\bar{\mathbf{D}} \mathbf{X} \mathbf{w}\|_2^2 \geq 0$$

since obviously $\mathbf{D} = \bar{\mathbf{D}}^\top \bar{\mathbf{D}}$.

$\Rightarrow \nabla_{\boldsymbol{\theta}}^2 \mathcal{R}_{\text{emp}}$ is positive semi-definite $\Rightarrow \mathcal{R}_{\text{emp}}$ is convex.

